A Listwise Approach to Coreference Resolution in Multiple Languages

Oanh Thi Tran, Bach Xuan Ngo, Minh Le Nguyen, and Akira Shimazu

School of Information Science, JAIST,
1 - 8 Asahidai, Nomi, Ishikawa, Japan 923 - 1211
{oanhtt, bachnx, nguyenml, shimazu}@jaist.ac.jp

Abstract. This paper presents a listwise approach as an alternative to commonly used pairwise approaches to the task of coreference resolution in multiple languages. In this listwise approach, all antecedent candidates are examined simultaneously and assigned corresponding scores expressing the probability that each candidate is coreferent with a given mention. The experimental results on the corpora of SemEval-2010 shared task 1 showed that our proposed system gave the good results in English and Spanish, and comparative results in Catalan when compared to previous participating systems. These results prove that this approach is appropriate and quite efficient for Coreference Resolution in Multiple Languages.

Keywords: listwise approach, coreference resolution, multiple languages.

1 Introduction

Reference resolution (Jurafsky and Martin, 2009) (chapter 21, section 21.4) is a task of determining to which entities are referred by which linguistic expressions. This task plays an important role in a large number of NLP applications such as Information Retrieval, Question Answering and Machine Translation. Therefore, it has attracted many attentions within the NLP community. Many works on various aspects (linguistic features (Ng, V., 2007), (Haghighi and Klein, 2009); machine learning models (Soon *et al.*, 2001); multiple languages (Recasens *et al.*, 2010a); and so on) of the coreference resolution task have been published.

Until the release of the SemEval-2010 task 1 (Recasens *et al.*, 2010a), there has no competition or public corpus that allows evaluating different coreference resolution systems in multiple languages. Most published systems only focus on a specific language and use the same data sets for example ACE or MUC corpora to train and test the systems. This makes the systems easy to unintentionally adapt themselves to the corpus but not to the problem in general. Therefore, this SemEval-2010 task 1 (Recasens *et al.*, 2010a) made it possible to evaluate and compare various automatic coreference resolution systems in the aspects of: (i) the portability of systems across languages, (ii) the relevance of different levels of linguistic information, and (iii) the behavior of scoring metrics.

This shared task attracted lots of researchers' attentions, but finally only six teams submitted their final results. The participating systems differed in terms of architecture, machine learning methods, etc. These systems mostly based on pairwise models, graph partitioning and entity-mention models. Unfortunately, these models suffered from an important weakness (Ng, V., 2010). In these models, each antecedent candidate is resolved independently with the other candidates. So the models could not determine the best candidate in the relation with the other candidates. To address this drawback, ranking models were proved to be a useful solution (Denis and Baldridge, 2007), (Ng, V., 2005), (Yang *et al.*, 2003). Motivated from ranking models, in this paper, we present our proposal approach for learning-based reference resolution task in multiple languages.

We exploit the listwise approach, which is originally proposed for learning to rank task in information retrieval (Cao *et al.*, 2007), to solve the SemEval-2010 Task on Coreference Resolution in Multiple Languages. This method allows the system to choose the best candidate for a given mention in the relation with other candidates. This means that all candidates will be examined simultaneously and the candidate with the highest score will be selected as a correct antecedent. This listwise approach has been successfully applied to information retrieval task (Cao *et al.*, 2007). Our experimental results on the corpora of SemEval-2010 shared task 1 showed that when applied to coreference resolution task, this new listwise approach usually gave the better results than previous approaches. When estimated on the latest metric BLANC, our proposed system got the state-of-the-art performance.

The rest of the paper is organized as follows. Section 2 reviews related work proposed for this shared task. Section 3 describes our listwise approach to this shared task. Section 4 presents experimental results on the corpora of this SemEval-2010 shared task. Finally, section 5 gives some conclusion and future work.

2 Related Work

In this section, we preview previous approaches of the systems participated in the SemEval-2010 shared task 1. The experimental results of these systems are also used to make an experimental comparison with our proposed approach's results. Here, we preview four systems: (1) RelaxCor system (Sapena *et al.*, 2010); (2) SUCRE system (Kobdani and Schutze, 2010); (3) TANL-1 system (Attardi *et al.*, 2010); and (4) UBIU system (Zhekova and Kubler, 2010). Table 1 presents an overview of the systems, their architecture and machine learning methods.

Table 1: Main characteristics of the previous systems.

Systems	System Architecture	Machine learning methods
RelaxCor	Graph Partitioning (solved by relaxation labeling)	Decision trees, Rules
SUCRE	Best-first clustering, Relational database model, Regular feature definition language	Decision trees, Nave Bayes, SVM, MaxEnt
TANL-1	Highest entity-mention similarity	MaxEnt
UBIU	Pairwise model	MBL

2.1 RelaxCor system

RelaxCor (Sapena *et al.*, 2010) is a constraint-based graph partitioning approach to coreference resolution solved by relaxation labeling. The approach combines the strengths of groupwise classifiers and chain formation methods in one global method. This system includes three phases:

Phase 1: Graph representation

Let $G = G(V, E)$ be an undirected graph. Each mention m_i in the document is presented as a vertex $v_i \in V$ in G. An edge $e_{ij} \in E$ is added to the graph for pairs of vertices (v_i, v_j) representing the possibility that both mentions corefer. A subset of constraints $C_{ij} \in C$ is used to compute the weight value w_{ij} of the edge connecting v_i and v_j.

Phase 2: Training process

Each mention pair (m_i, m_j) in training document is evaluated by the set of feature functions which form a positive example if the mention pair corefers, and a negative otherwise. For each type of mention mj (for example: pronoun, named entity or nominal), a decision tree is generated and a set of rules is extracted with C4.5 rule-learning algorithm.

Given the training corpus, the weight of a constraint C_k is related with the number of examples where the constraint applies and how many of them corefer.

Phase 3: Resolution Algorithm

The algorithm solves weighted constraint satisfaction problem dealing with the edge weights w_{ij}. In this manner, each vertex is assigned to a partition satisfying as many constraints as possible. The algorithm assigns a probability for each possible label of each variable (corresponding to each vertex in G). The process updates the weights of the labels in each step until convergence. Finally, the assigned label for a variable is the one with the highest weight.

2.2 SUCRE system

This system developed a feature engineering which can help reducing the implementation effort for feature extraction. SUCRE has a novel approach to model an unstructured text corpus in a structured framework by using a relational database model and a regular feature definition language to define and extract the features.

In learning, there are four classifiers integrated in SUCRE: Decision tree, Nave bayes, Support vector machine and maximum entropy. However, finally the best reported results were achieved with Decision tree. In decoding, the coreference chains are created. The system uses best-first clustering. It searches for the best predicted antecedent from right-to-left starting from the end of the document.

2.3 TANL-1 system

The system is built based on highest entity-mention similarity. The authors applied Maximum Entropy classifier to determine whether two mentions refer to the same entity. The classifier is trained using the features extracted for each pair of mentions. If the pairwise classifier assigns a probability greater than a given threshold to the fact that a new mention belongs to a previously identified entity, it is assigned to that entity. In the case that more than one entity has a probability greater than the threshold; the mention is assigned to the one with highest probability by using best-first clustering strategy.

2.4 UBIU system

Classification in UBUI is based on mention pairs. The UBIU system used a combination of machine learning, in the form of memory-based learning (MBL) in the implementation of TiMBL (Daelemans *et al.*, 2007), and language independent features. MBL uses a similarity metric to find the k nearest neighbors in the training data in order to classify a new example.

Despite of the difference in feature engineering, learning methods and some processing techniques, it can be seen that three later systems - SUCRE, TANL-1, and UBIU - belong to the approach called pairwise approach. The typical machine learning approach of these three systems includes two steps:

- Classification: systems evaluate whether each pair of mentions is coreferent with each other or not.
- Formation coreference chain: Given the previous classification, the systems form coreference chain (mostly based on best-first clustering).

The approach presented in RelaxCor system joined classification and chain formation into the same step. In this manner, decisions are taken considering the whole set of mentions, ensuring consistency and avoiding that classification decisions are independently taken.

3 A listwise approach to coreference resolution task

3.1 Coreference resolution as a ranking problem

In previous models, a classifier is trained to determine whether two NPs are coreferent or not. Instances are created based on mention to be resolved and an antecedent candidate. However, those models suffer from an important weakness. Since each antecedent candidate for an anaphoric NP is considered independently of the others, it cannot determine how good an antecedent candidate is relative to other candidates. To address this drawback, ranking models were proved to be useful solutions (Denis and Baldridge, 2007), (Ng, V., 2005), (Yang *et al.*, 2003). In the ranking model, most authors use tournament by (Iida, 2003) and twin-candidate model by (Yang *et al.*, 2003) and cluster-ranking by (Rahman and Vincent, 2009) to solve the problem of ranking antecedent candidates. However, this weakness is not fully solved in the aspect that these models cannot examine all antecedent candidates at the same time. They only directly compare pairs of antecedent candidates by building a preference classifier based on triples of NP mentions.

In the next sub-section, we will present a new listwise approach - ListNet method - to this task in multiple languages. It addresses the drawback of previous approaches to this coreference resolution task as discussed above.

3.2 ListNet method

This sub-section briefly presents ListNet method - a listwise approach with Neural Network as the model and Gradient Descent as the optimization algorithm. This method was proposed by (Cao *et al.*, 2007) for the task of learning to rank. We first state the learning problem in listwise approach to learning to rank task. Then, we present ListNet method and the learning algorithm of ListNet. In the following description, we use superscript to denote the id of the mention to be resolve and subscript to denote the id of a candidate in the antecedent candidate list.

In listwise approach to learning to rank, a set of m samples $S = s^{(1)}, s^{(2)}, \dots, s^{(m)}$ is given. Each sample $s^{(i)}$ consists of an object list $o^{(i)} = o_1^{(i)}, o_2^{(i)}, \dots, o_{n^{(i)}}^{(i)}$, where $o_j^{(i)}$ denotes the j^{th} object and $n^{(i)}$ denotes the number of objects in i^{th} sample. Furthermore, each object list $o^{(i)}$ is associated with a list of scores $y^{(i)} = y_1^{(i)}, y_2^{(i)}, \dots, y_{n^{(i)}}^{(i)}$, where $y_j^{(i)}$, a real number, is the score of the object $o_j^{(i)}$. In coreference resolution task, for example, a sample $s^{(i)}$ is associated with a mention $m^{(i)}$to be resolved, each object $o_j^{(i)}$ corresponds to an antecedent candidate $c_j^{(i)}$, and score $y_j^{(i)}$ denotes the judgment on an antecedent candidate $c_j^{(i)}$ with respect to the mention $m^{(i)}$ (the value of $y_j^{(i)}$ expresses how relevant coreference an antecedent candidate $c_j^{(i)}$ is with a mention $m^{(i)}$ to be resolved).

A feature function ϕ will produce a real-valued feature vector for each object $x_j^{(i)} = \phi(o_j^{(i)})$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n^{(i)}$. A list of feature vectors $x^{(i)} = x_1^{(i)}, x_2^{(i)}, \dots, x_{n^{(i)}}^{i}$ and the corresponding list of scores $y^{(i)} = y_1^{(i)}, y_2^{(i)}, \dots, y_{n^{(i)}}^{(i)}$ will form a training instance $(x^{(i)}, y^{(i)})$. The training set can be represented by the following set: $D = (x^{(i)}, y^{(i)})_{i=1}^{m}$.

In training phase, we want to learn a ranking function f, that produces a real-valued score $f(x_j^{(i)})$ for each feature vector $x_j^{(i)}$.

Suppose that $z^{(i)} = \Big(f(x_1^{(i)}), f(x_2^{(i)}), \dots, f(x_{n^{(i)}}^{(i)})\Big)$ is the list of scores produced by f on a list of feature vectors $x^{(i)} = x_1^{(i)}, x_2^{(i)}, \dots, x_{n^{(i)}}^{(i)}$, and L is a loss function defined on two lists of scores $y^{(i)}$ and $z^{(i)}$. We want to minimize the total losses on the training data:

$$\sum_{i=1}^{m} L(y^{(i)}, z^{(i)}) \tag{1}$$

In ranking phase, given a new sample $s^{'}$ (a list of new objects $o^{'}$), we first construct a list of feature vectors $x^{'}$ using feature function ϕ, and then produce a list of scores $y^{'}$ using ranking function f. Finally, objects are ranked in descending order of the scores.

ListNet is a listwise method for learning to rank task. ListNet uses Cross Entropy metric as loss function, Neural Network as model, and Gradient Descent as learning algorithm. If we use a linear Neural Network model, the score of a feature vector can be calculated as follows:

$$f_\omega(x_j^{(i)}) = \left\langle \omega, x_j^{(i)} \right\rangle \tag{2}$$

where $\langle ., . \rangle$ denotes an inner product.

The following algorithm 1 shows learning steps of this ListNet method.

Algorithm 1 Learning Algorithm of ListNet method (cited from the paper of (Cao *et al.*, 2007))

```
Input: Set of training instances: $(x^{(i)}, y^{(i)})_{i=1}^{m}$
Parameter: iteration number $T$ and learning rate $\eta$
Initialize parameter $\omega$
  for $t = 1 \rightarrow T$ do
    for $i = 1 \rightarrow m$ do
      Input $x^{(i)}$ to Neural Network and Compute score list $z^{(i)}(f_\omega)$ with current value of $\omega$
      $z^{(i)}(f_\omega) = \left( f_\omega(x_1^{(i)}), \ldots, f_\omega(x_{n^{(i)}}^{(i)}) \right)$
      Compute gradient $\Delta\omega$ using equation (3)
      Update $\omega = \omega - \eta \times \Delta\omega$
    end for
  end for
Output: Neural Network model $\omega$
```

$\Delta\omega$ is computed using the following formula:

$$\Delta\omega = \frac{\delta L(y^{(i)}, z^{(i)}(f_\omega))}{\delta\omega} =$$

$$- \frac{1}{\sum_{j=1}^{n^{(i)}} exp(y_j^{(i)})} \sum_{j=1}^{n^{(i)}} exp(y_j^{(i)}) \frac{\delta f_\omega(x_j^{(i)})}{\delta\omega} + \frac{1}{\sum_{j=1}^{n^{(i)}} exp(f_\omega(x_j^{(i)}))} \sum_{j=1}^{n^{(i)}} exp(f_\omega(x_j^{(i)})) \frac{\delta f_\omega(x_j^{(i)})}{\delta\omega} \tag{3}$$

3.3 Modeling the listwise approach to coreference resolution task

In previous models, they only determine how good a candidate antecedent is relative to the anaphoric NP, but not how good a candidate antecedent is relative to other candidates. In other words, they fail to answer the question of which candidate antecedent is the most probable. Our proposed model will allow us to determine which candidate antecedent is the most probable given an NP to be resolved. For example, we have four mentions named A, B, C, and D in the order of their occurrence in the document. Given that we are resolving an anaphoric mention D to determine a true antecedent among A, B, and C.

The following describes our training and resolution phase of the system.

Training phase

In this phase, trained instance is created as follows:

The training instances for this listwise approach are built based on a mention to be resolved D and a list of its antecedent candidates together with their scores. The candidate set includes all mentions occurring before an anaphoric mention D. The score denotes whether each candidate corefers with mention D and whether each candidate is closest to mention D if they corefer. The way of getting our training data is a way to rank candidates based on coreference and distance criterion. This way is somewhat like a way of human score. In learning, the system has to induce the model that determines these rankings based on not only the distance but also other criteria such as other features and relations between them. In the above example, if we create an instance corresponding to mention D, we have an instance in the form of $(D - A : Pr(D - A); D - B : Pr(D - B); D - C : Pr(D - C))$. In that $Pr(D - x)$ is determined as follows:

$$Pr(D-x) = \begin{cases} 0 & \text{if } D \text{ is not coreferent with } x \\ 1 & \text{if } D \text{ is coreferent with } x \text{ and } x \text{ is closest to } D \\ 0.5 & \text{if } D \text{ is coreferent with } x \text{ and } x \text{ is not closest to } D \end{cases}$$

These training instances are used to learn parameters of the Neural Network model ω according to the algorithm 1.

Resolution phase

Figure 1 visualizes the resolution phase. In our listwise approach, we create instance as a list of candidates in learning and the learning function will assign a score for each candidate in that list. The candidate is chosen as a true antecedent is the one that has the highest score. To allow a mention to be non-anaphoric, we set up a threshold θ to determine whether a given mention is anaphoric or not. This parameter will be chosen using the development set. In figure 1, we assume that B is selected as an exact antecedent of mention M among the candidate list of A, B, C, and so on.

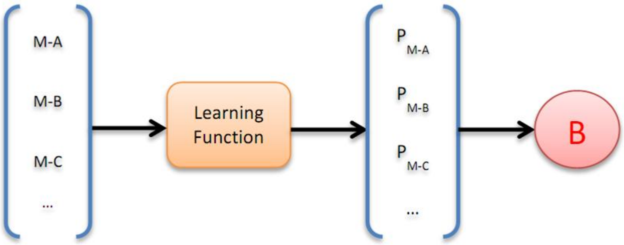

Figure 1: Visualizing the resolution phase.

4 Experiments

4.1 Describing the corpus and evaluation metrics

In these experiments, we used the corpora of SemEval-2010 task 1 on Coreference Resolution in Multiple Languages. We tested our system using three different languages (Catalan, English, and Spanish). The size of the task datasets are provided in the table 2:

Table 2: Size of the task datasets.

Languages	Training			Development			Testing		
	#docs	#sents	#tokens	#docs	#sents	#tokens	#docs	#sents	#tokens
Catalan	829	8,709	253,513	142	1,445	42,072	167	1,698	49,260
English	229	3,648	79,060	39	741	17,044	85	1,141	24,206
Spanish	875	9,022	284,179	140	1,419	44,460	168	1,705	51,040

In the experiments, we evaluate our system using closed gold-standard setting. It means that we use the gold-standard columns with true mention boundaries and our system was built strictly with the information provided in the task datasets. This is because our system focuses on evaluating various approaches of previous participating systems versus our proposed listwise approach.

To evaluate our system, we also use four metrics which are CEAF (Luo, 2005), MUC (Vilain *et al.*, 1995), BCUB (Bagga and Baldwin, 1998) and BLANC scores (Recasens and Marti, 2010b) provided by this shared task. The first three measures have been widely used, while BLANC is a proposal of a new measure interesting to test.

MUC-6/7 (Vilain *et al.*, 1995)

This is the oldest and most widely-used metric measure. This metric is based on coreference links. First, we count the number of common links between the reference (or ”truth”) and the system output (or ”response”). The link precision is the number of common links divided by the number of links in the system output, and the link recall is the number of common links divided by the number of links in the reference.

BCUB (Bagga and Baldwin, 1998)

The MUC metric yields unintuitive results because of two main shortcomings. First, it does not give any credit for single-mention entities since no link can be found in these entities. Second, all errors are considered to be equal because in some tasks, some coreference errors do more damage than others. These drawbacks lead to the proposal of BCUB metric. This metric first computes a precision and recall for each individual mention, and then takes the weighted sum of these individual precisions and recalls as the final metric. The choice of the weighting scheme is determined by the task for which the algorithm is going to be used.

CEAF (Luo, 2005)

The BCUB metrics still has its own problems: for example, the mention precision/recall is computed by comparing entities containing the mention and therefore an entity can be used more than once. Thus, they proposed Constrained Entity-Aligned F-measure or CEAF metric. It finds the best one-to-one mapping entities between the subsets of reference and system entities. They are aligned by maximizing the total entity similarity under the constraint that a reference entity is aligned with at most one system entity, and vice versa. After that, it computes the recall, precision and F-measure.

BLANC (Recasens and Marti, 2010b)

BLANC, a measure obtained by applying the Rand index (Rand 1971) to coreference resolution and taking into account the shortcomings of above previous metrics. The Rand index seems to be especially adequate for evaluating coreference since it allows us to measure ’non-coreference’ as well as coreference links. Despite of its shortcomings, it addresses to some degree the drawbacks of previous metrics.

4.2 Feature Sets

In this task, the feature sets were selected from the feature pool presented in (Haghighi and Klein, 2009). We selected 22 features which are divided into 3 groups as described in more detail in table 3. These features are popular and available in all languages of this SemEval-2010 shared task 1.

4.3 Experimental results

By implementing ListNet, our system had to choose the set of three parameters which are (1) number of iteration T; (2) learning rate η; and (3) the threshold θ to determine a candidate is coreferent with a given mention or not. To determine the best parameter set, we varied their values and selected parameters that maximize the sum of four metrics based on the development sets. After that, we used these parameters to evaluate our proposed system on the test sets. The best parameter sets for three languages are presented in the table 4.

The experimental results are compared with four models which are (1) RelaxCor system (Sapena *et al.*, 2010); (2) SUCRE system (Kobdani and Schutze, 2010); (3) TANL-1 system (Attardi *et al.*, 2010); and (4) UBIU system (Zhekova and Kubler, 2010). Table 4 shows our experimental results of the proposed model for three languages using four metrics.

For Catalan language, our system got the best result on BLANC F-scores. It beat TANL-1 and UBIU systems in all four F-scores. In comparison with RelaxCor system, MUC and BLANC F-score increase significantly from 42.5 to 55.42 and from 59.7 to 67.13; CEAF and BCUB F-score decrease from 70.5 to 67.15, and from 79.9 to 76.35. In comparison with SUCRE system, our system only increases on BLANC F-scores and decreases on three remaining F-scores.

Table 3: The feature set for all languages.

Features describing m_j, a candidate antecedent	
1. PRONOUN_1	Y if m_j is a pronoun; else N
2. SUBJECT_1	Y if m_j is a subject; else N
3. NESTED_1	Y if m_j is a nested NP; else N
Features describing m_k, the mention to be resolved	
4. NUMBER_2	SINGULAR or PLURAL, determined using a lexicon
5. GENDER_2	MALE, FEMALE or UNKNOWN, determined using a list of common first names
6. PRONOUN_2	Y if m_k is a pronoun; else N
7. NESTED_2	Y if m_k is a nested NP; else N
8. SEMCLASS_2	The semantic class of m_k
9. HEAD_MATCH	C if the mentions have the same head noun; else I
10. STR_MATCH	C if the mentions are the same string; else I
11. SUBSTR_MATCH	C if one mention is a substring of the other; else I
12. NUMBER	C if the mentions agree in number; I if disagree; NA if numbers for one or both mentions cannot be determined
13. GENDER	C if the mentions agree in gender; I if disagree; NA if genders for one or both mentions cannot be determined
14. AGREEMENT	C if the mentions agree in both gender and number; I if they disagree in both number and gender; else NA
15. BOTH_PRONOUNS	C if both mentions are pronouns; I if neither are pronouns; else NA
16. SEMCLASS	C if the mentions have the same semantic class; I if they don't; NA if the semantic class information for one or both mentions cannot be determined
17. DISTANCE	Binned values for sentence distance between the mentions
Additional features describing the relationship between m_j, m_k	
18. NUMBER'	The concatenation of the NUMBER_2 feature values of m_j and m_k
19. GENDER'	The concatenation of the GENDER_2 feature values of m_j and m_k
20. PRONOUNS'	The concatenation of the PRONOUN_2 feature values of m_j and m_k
21. NESTED'	The concatenation of the NESTED_2 feature values of m_j and m_k
22. SEMCLASS'	The concatenation of the SEMCLASS_2 feature values of m_j and m_k

For English language, our system got the best results on three F-scores of CEAF, BCUB and BLANC and got the second best on the remaining MUC F-score. In that, CEAF and BLANC F-scores increase significantly from the previous highest scores 75.6 to 78.58, and from 70.8 to 75.66. Our system also overcomes three systems of RelaxCor, TANL-1 and UBIU in all four metrics.

For Spanish language, our system got the best results on two F-scores of BLANC and MUC. For the two remaining F-scores of CEAF and BCUB, we got the results which are comparative to previous best scores (CEAF: 69.15 in comparison with 69.8; and 77.81 in comparison with 78.2). Our system beat TANL-1 and UBIU system. Compared with RelaxCor, our system got higher results on three F-scores of CEAF (from 66.6 to 69.15), MUC (from 24.7 to 57.82) and BLANC (from 55.6 to 67.38) which are all significant increase. For the remaining BCUB F-score, our system decreases insignificantly (from 78.2 to 77.81). Compared to SUCRE system, our system increases the F-score of MUC, BCUB and BLANC and decrease insignificantly on the CEAF F-score.

4.4 Discussion

The experimental results on these corpora showed that our proposed system gave the good results in English and Spanish; and comparative results in Catalan when compared to previous participating systems. When applied to coreference resolution task in multiple languages, this new listwise approach overcomes most of previous approaches for all four available metrics. For other systems which our system cannot overcome, usually we got the comparative results or insignificantly decrease in one metric and significantly increase in other remaining metrics.

Among lots of metrics proposed for evaluating a coreference resolution system, none of them is fully adequate. Each metrics has its own strong points as well as weak points as we discussed

Table 4: Experimental results of the proposed model for three languages and four metrics.

Languages	Systems	MUC			BCUB			CEAF			BLANC		
		R	P	F1	R	P	F1	R	P	F1	R	P	F1
English	RelaxCor	21.9	72.4	33.7	74.8	97.0	84.5	75.6	75.6	75.6	57.0	83.4	61.3
	SUCRE	68.1	54.9	**60.8**	86.7	78.5	82.4	74.3	74.3	74.3	77.3	67.0	70.8
	TANL-1	23.7	24.4	24.0	74.6	72.1	73.4	75.0	61.4	67.6	51.8	68.8	52.1
	UBIU	17.2	25.5	20.5	67.8	83.5	74.8	63.4	68.2	65.7	52.6	60.8	54.0
	Our system 2 - 0.005 - 0.25	48.62	62.4	54.66	81.29	89.19	**85.05**	78.17	78.99	**78.58**	73.75	77.92	**75.66**
Spanish	RelaxCor	14.8	73.8	24.7	65.3	97.5	**78.2**	66.6	66.6	66.6	53.4	81.8	55.6
	SUCRE	52.7	58.3	55.3	75.8	79.0	77.4	69.8	69.8	**69.8**	67.3	62.5	64.5
	TANL-1	16.6	56.5	25.7	65.2	93.4	76.8	66.9	64.7	65.8	52.5	79.0	54.1
	UBIU	9.6	18.8	12.7	46.8	77.1	58.3	45.7	59.6	51.7	52.9	63.9	54.3
	Our system 5 - 0.01 - 0.25	58.15	57.49	**57.82**	78.5	75.9	77.81	69.13	69.16	69.15	71.69	64.62	**67.38**
Catalan	RelaxCor	29.3	77.3	42.5	68.6	95.8	**79.9**	70.5	70.5	**70.5**	56.0	81.8	59.7
	SUCRE	51.4	58.4	**56.2**	76.6	77.4	77.0	68.7	68.7	68.7	72.4	60.2	63.6
	TANL-1	17.2	57.7	26.5	64.4	93.3	76.2	66.0	63.9	64.9	52.8	79.8	54.4
	UBIU	8.8	17.1	11.7	47.8	76.3	58.8	46.6	59.6	52.3	51.6	57.9	52.2
	Our system 2 - 0.001 - 0.15	55.28	55.56	55.42	77.11	75.6	76.35	67.13	67.16	67.15	70.79	64.67	**67.13**

in section 4.1. This situation makes it hard to successfully compare systems. Getting the state-of-the-art performance based on all this four common metrics seems to be a difficult task. Until now there is no common agreement on a standard measure for coreference resolution task.

However, based on formulas and characteristics of each metric, we saw that later-proposed metrics usually give the better quality than metrics proposed early. If using this criterion, we saw that we got the highest F-score for the latest proposed BLANC metric in all three languages. In other words, our proposed system got the state-of-the-art performance for coreference resolution task in multiple languages.

5 Conclusion and Future Work

In this paper, we present a new listwise approach to the task of SemEval-2010 task 1 on coreference resolution in multiple languages. This new listwise approach allows all candidate antecedents are considered simultaneously and therefore brings in more benefit than traditional pairwise approaches. The experimental results on the public corpora showed that this new proposed approach gave relatively good performance in all three languages. For the latest proposed metric BLANC, we got the state-of-the-art performance for this coreference resolution task in multiple languages.

For the future works, we will continue to do experiments using other settings of the SemEval-2010 task 1 to fortify the strength of this listwise approach.

A straightforward generalization of twin-candidate model is the ranker model proposed by (Denis and Baldridge, 2007). In this ranker model, the computation of the model's expectation of a feature is directly based on the probabilities assigned to the different candidates by using supervised maximum entropy ranking approach. In the future, we also would like to investigate other ranker models like this one on the corpora of this SemEval-2010 shared task number 1.

Acknowledgments

This work was partly supported by the 21st Century COE program 'Verifiable and Evolvable e-Society', Grant-in-Aid for Scientific Research, Education and Research Center for Trustworthy e-Society.

References

Attardi, G., Rossi, S. D., and Simi, M. 2010. TANL-1: coreference resolution by parse analysis and similarity clustering. *SemEval-2*, pp.108-111.

Bagga, A. and Baldwin, B. 1998. Algorithms for scoring coreference chains. *LREC Workshop on Linguistic coreference*, pp.563-566.

Cao Z., Qin T., Liu T.Y., Tsai M.F., Li H. 2007. Learning to rank: from pairwise approach to listwise approach. *ICML*, pp. 129-136.

Daelemans, W., Zavrel, J., Sloot, K., and Bosch, A. 2007. TiMBL: Tilburg memory based learner version 6.1 reference guide. *Technical Report ILK 07-07, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University*.

Denis, P. and Baldridge, J. 2007. A ranking approach to pronoun resolution. *International Conference on Artificial Intelligence*, pp.1588-1593.

Haghighi, A., Klein, D. 2009. Simple Coreference Resolution with rich syntactic and semantic features. *Empirical methods in Natural Language Processing*, pp.1152-1161.

Iida, R., Inui, K., Takamura, H., and Matsumoto, Y. 2003. Incorporating contextual cues in trainable models for coreference resolution. *EACL Workshop on the computational Treatment of anaphora*, pp.23-30.

Jurafsky D., Martin J.H. 2009. Speech and Language Processing. *Prentice Hall Series in Artificial Intelligence*, 2nd Edition.

Kobdani, H. and Schutze, H. 2010. SUCRE: Modular system for coreference resolution. *SemEval-2*, pp.92-95.

Luo, X. 2005. On coreference resolution performance metrics. *HLT-EMNLP*, pp.25-32.

Ng, V. 2005. Supervised ranking for pronoun resolution: Some recent improvements. *AAAI*, pp.1081-1086.

Ng, V. 2007. Semantic class induction and co-reference resolution. *ACL*, pp.536-543.

Ng, V. 2010. Supervised Noun phrase coreference research: The first fifteen years. *Annual Meeting of the Association for Computational Linguistics*, pp.1396-1411.

Rahman, A. and Vincent Ng. 2009. Supervised models for coreference resolution. *Empirical Methods in Natural Language Processing*, pp.968-977.

Recasens, M., Marquez, L., Sapena, L., Marti, M., Taule, M., Hoste, V., Poesio, M., Versley, Y. 2010a. SemEval-2010 Task 1: Co-reference Resolution in Multiple Languages. *International Workshop on Semantic Evaluation, ACL*, pp.1-8.

Recasens, M. and Hovy, E. 2010b. BLANC: Implementing the Rand Index for coreference evaluation. *In prep.*

Soon W.M., Ng H.T., Lim D.C.Y 2001. A machine learning approach to co-reference resolution of noun phrases. *Computational Linguistics*, pp.521-544.

Sapena, E., Padr, L., and Turmo, J. 2010. RelaxCor: A Global relaxation labeling approach to coreference resolution for the SemEval-2 Coreference Task. *SemEval-2*, pp.88-91.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. 1995. A model-theoretic coreference scoring scheme. *MUC-6*, pp.45-52.

Yang, X., Zhou, G., Su, J. and Tan, C.L. 2003. Coreference resolution using competitive learning approach. *ACL*, pp.176-183.

Zhekova, D., and Kubler, S. 2010. UBIU: A language-independent system for coreference resolution. *SemEval-2*, pp.96-99.